Anna Strasser (Ed.)

Anna‘s AI Anthology

How to live with smart machines?

MoMo berlin Philosophische KonTexte Band 9

Anna Strasser (ed.)

Anna's AI Anthology

Anna Strasser (ed.)

Anna's AI Anthology

How to live with smart machines?

MoMo Berlin • Philosophische KonTexte Vol. 9

xenomoi

Imprint

Anna Strasser (Ed.)

Anna's AI Anthology
How to live with smart machines?

ISBN:
Hardcover: 978-3-942106-90-0
Paperback: 978-3-942106-98-6
PDF-Version: 978-3-942106-99-3

© 2024 xenomoi Verlag, Berlin
Set in Palatino Linotype 10 Pt.

All drawings including the supplementary graphic novel:
Moritz Strasser, Freiburg i.Br.

Cover design, Layout and production:
xenomoi Verlag e.K., Heinersdorfer Str. 16, D - 12209 Berlin
Phone.: 030 - 755 11 712 ▪ www.xenomoi.de ▪ info@xenomoi.de

This publication including its supplementary graphic novel is in copyright. Subject to statutory exception and to the provisions of relevant collective licensing agreements, no reproduction of any part and in any form may take place without the written permission of xenomoi Verlag.

First published 2024

Bibliographic information of the German National Library:
The German National Library lists this publication in the German National Bibliography; detailed data is available on the Internet at https://portal.dnb.de/opac.htm.

Contents

Preface by the publisher 3
Introduction by the editor 7

• • •

Daniel Dennett (Tufts University, Santa Fe Institute)
We are all cherry-pickers 15

Alessio Tacca & Frederic Gilbert (University of Tasmania)
Just copy-paste me! Assessing the risks of epistemic dependence on Large Language Models 31

• • •

Keith Frankish (University of Sheffield)
What are large language models doing? 53

Paula Droege (Pennsylvania State University)
Full of sound and fury, signifying nothing 79

Joshua Rust (Stetson University)
Minimal agency in living and artificial systems 93

• • •

Stephen Butterfill (Warwick University)
What mindreading reveals about the mental lives of machines 123

Ying-Tung Lin (National Yang Ming Chiao Tung University)
The fluidity of human mental attribution to large language models 149

Sven Nyholm (LMU Munich)
Generative AI's gappiness: Meaningfulness, authorship, and the credit-blame asymmetry 167

• • •

Anna Strasser (DenkWerkstatt Berlin, LMU Munich) & Eric Schwitzgebel (UC Riverside)
Quasi-sociality: Toward asymmetric joint actions 195

Syed AbuMusab (University of Kansas)
Large language models belong in our social ontology 213

Constant Bonard (University of Bern)
Can AI and humans genuinely communicate? 243

• • •

Henry Shevlin (Leverhulme Centre for the Future of Intelligence, University of Cambridge)
Consciousness, machines, and moral status 277

Michael Wilby (Anglia Ruskin University) & Anna Strasser (DenkWerkstatt Berlin, LMU Munich)
Situating machines within normative practices: Agency, moral responsibility & the AI-Stance 309

Appendix:

The Human & The AI
A graphic novel by Anna Strasser and Moritz Strasser

Wolfgang Sohst

Foreword

It was a great pleasure for my small academic publishing house, *xenomoi*, to be involved in the creation of an anthology that brings together philosophical perspectives on how recent advances in artificial intelligence might affect humanity.

In historical terms, the emergence of digital computers and artificial neural networks is a recent phenomenon. Due to recent advances in generative AI, it seems as if cultural evolution has reached an important turning point, as it is now possible to train these machines to produce linguistic output that makes sense to humans – output that *might* have semantic content. The question arises as to whether the proud designers have come close to the Promethean goal of breathing the fire of human intelligence into inanimate nature. How, then, should we – the 'real' humans – deal with the influence of this new technology, which helps us and, at the same time, competes with us?

It is not surprising that professional philosophers have weighed in on these important questions, as can be seen from the flood of recent publications. Anna Strasser, however, has had an even more ambitious idea. This book does not follow the usual custom in the world of academic books, where the experts dispense their wisdom to the listening public like a precious summer rain. Rather, Anna has constructed her book in such a way as to create a lively, bidirectional contact between its contributors and its readers. After all, one of the obvious dangers of the emergence of machines that seem to interact semantically with us is that they may be perceived as psychologically structured and socially acting beings. There is a risk that they will be mistaken for persons and thus penetrate the fabric of social existence through an almost undetectable deception, in comparison with which Odysseus' Trojan horse was a harmless toy. It is already impossible to distinguish with certainty between machine-generated text and human-created text. Anna has tried to mitigate this risk by producing a concrete book that, while discussing this indistinguishability, makes the human authorship of its own texts clearly recognizable.

Following this *leitmotif*, Anna organized the production of the texts in a novel way. She asked the contributing authors to exchange their drafts with a team partner and to comment on each other's work, allowing them

to reassure each other of their human authorship. This procedure also created a dialogical moment, which posed a challenge of its own kind for the respective authors, who were confronted with the thoughts of other contributors who had different priorities and came to different conclusions. By placing themselves in this tension before the publication of their contributions, the authors' philosophical writing was brought down from the lonely heights of their individual thinking to the social level. In addition, the editor provided extensive feedback to each of the contributors, which enabled them to make further cross-references to the other contributions.

But this volume does even more than that. By introducing each contribution with a page that provides two portraits of the respective author (a photograph and a sketch from the graphic novel) plus some personal information and a QR code that can be used to obtain more details about the person behind the text, it brings the philosophical discussion of socially significant questions back into the sphere of our everyday human coexistence. This is intended to make it clear that these are not machine-generated texts.

Moreover, in order to emphasize the individuality of the authors and their contributions, which fundamentally distinguishes them from abstract language models, Anna has deliberately refrained from copyediting the contributions to make them conform to a specific variant of English, such as US or British. In this way, each author is able to speak in their own voice.

Going even further, Anna came up with the idea of supplementing the anthology with a graphic novel, which she and her brother Moritz created especially for this book, in order to add a vivid pictorial and emotional quality that philosophical texts alone cannot achieve. A special thing about this graphic novel is that the protagonist is himself a reader of the very anthology that you are currently holding in your hands. In this way, it becomes clear that this book aims to stimulate an active, albeit virtual, interaction between the reader and the authors, which functions quite differently from pseudo-social human-machine interactions.

However, this unconventional idea posed some technical problems for me as the publisher. In the hardcover edition, I realized the most sophisticated variant by inserting the graphic novel as a separate booklet printed in full color at the back of the book. This, however, was not possible in the paperback and PDF versions. There, the graphic novel appears as a supplement to all the philosophical contributions at the end of the regular book. It is fair to say that large, established publishers of contemporary philosophy are not prepared to support such experiments, for economic

reasons at least. It was precisely the small size of *xenomoi* that enabled us to create a book that stands out from the crowd. We are more flexible when it comes to layout standards, which are elsewhere relatively rigid. And the hardcover version at least is made in-house, with our own hands and machinery, which makes an unconventional book format possible.

More than 150 years ago, the concept of *alienation* or *estrangement* (German: *Entfremdung*) was popularized through the writings of Karl Marx and Friedrich Engels. They promoted it as a warning against a social development in which the undeniable benefits of industrialization threatened to bring with them a drastic erosion of individual and collective quality of life. This book has no political or utopian ambitions. Nevertheless, through the way it was conceived and produced, it aims to counteract alienation in modern societies, where everything is drifting apart like the planks of a broken hull on the wild surface of the sea of social evolution. This book is an attempt to bring together what really belongs together – the meaningful interaction between people and their felt concerns, which gives us the freedom to shape our living together. These are truly great ambitions for a small book. But this book may at least give a small nudge in that direction. And like the proverbial flapping of a butterfly's wings, perhaps it will have a greater effect than might be expected.

Berlin, May 2024

Anna Strasser

Introduction

I start with a few words about the genesis of this anthology. We all remember the pandemic years, when Covid forced academia to devise online forms of interaction, as neither traveling nor group meetings were possible. At one of those virtual happenings, the poster session of CogSci 2021, I ended up talking to Eric Schwitzgebel and Mathew Crosby about the opportunities and challenges presented by the rapid development of generative AI technology, and, in particular, the latest generation of large language models (LLMs) (at the time, GPT-3). As we talked, the idea was born that we should team up to find out if it is possible to fine-tune such a large language model so that it could make serious philosophical statements. After a few pilots, I approached Daniel Dennett to ask him whether he would allow us to use his published work to fine-tune GPT-3 (Schwitzgebel et al., 2023; Strasser et al., 2023). To honor Dennett's role in this project, I quote from the account he gave in his autobiography:

> I have entered the fray thanks to Anna Strasser, a German philosopher who came to the Center for Cognitive Studies as a visiting fellow in 2018. She and philosopher Eric Schwitzgebel have, with my permission, trained GPT-3 on almost all my published works – well over a million words – to create DigiDan (Steve Barney's name for it), which (*not "who"!)* can answer interview questions addressed to me about my work and beliefs with uncanny verisimilitude and accuracy. Anna asked me ten questions about central ideas in my work, and I answered them – without any dissembling, unlike the confederates in year two of the Loebner Prize competition. Then she asked the same questions of DigiDan four times, and (without cherry-picking) took all five answers and put them to hundreds of judges online, who were asked to pick the genuine Dennett answer from the set of five and also to rank all five answers in how "like what Dennett would say" each answer was. She also invited several dozen "Dennett experts" (philosophers or cognitive scientists who have read more than a thousand pages of Dennett, or been colleagues or coauthors of mine) to take the test, and another group of philosophers and philosophy students.
>
> The results are surprising, at least at first glance. Even the "Dennett experts" were lured into choosing at least two DigiDan answers, averaging a little over 5 out of 10 right (which is still way above chance, given five choices; chance is about 1.2 right). The naïve judges were just about at chance, which was expected, since they were really just guessing. The more philosophy training you had, the better you did,

which is a relief for me to learn, and the mistaken choices by the experts were almost all DigiDan answers that I actually approved – though none of them quite won the hallowed "I wish I'd said that" award. There is much, much more I want to say about the DigiDan experiment, which in fact has inspired me to start writing a new paper on one of my long-term head-scratchers: How do we form speech acts? In *Consciousness Explained* I ventured some hunches about "how words do things to us" that contrasted" *bureaucratic* models, which have a Central Meaner – a kind of CEO – at the top, with *pandemonium* models, where many agents scramble and compete to get themselves into speech acts, but while I was all in favor of pandemonium, I could only wave my hands about how it might work. DigiDan outputs are well-formed English sentences, but it can be easily unmasked if you know what you're doing. If DigiDan doesn't understand anything (and it doesn't) it surely has a lot of competence without comprehension. Could the *sort* of competence it has also be the sort of competence some part of your brain has when it somehow contributes to framing an actual speech act, *expressing* a belief, *asking* a question, *entertaining* a hypothesis? GPT-3 has been dismissed by computational linguist Emily Bender as a "stochastic parrot," but in fact, the random candidates it generates are very nonrandomly filtered by all the patterns the system has detected and used to adjust its connections. Suppose it fed its uncomprehended outputs to a ... *cherry-picker* that could swiftly discard the junk and home in on a few more-promising candidates, and suppose that cherry-picker sent its winners to *another* cherry-picker which (not "who") could make a few mutations and pass the products on to yet another round of cherry-picking? In short, couldn't there be a Darwinian mutation-and-selection cascade that turned uncomprehending competence into (*sorta*) comprehending competence? And suppose all of this is largely unconscious processing – except when we're being particularly careful with our speech acts, deliberately checking them and rechecking them for "what we want to say."

DigiDan could already be used *by me* as a generator of possible ways of putting things that I could then edit and revise, confident that I would not be wasting my time on irrelevant babble because of its competence. The composer/computer programmer David Cope created his EMI (Experiments in Musical Intelligence) as a tool to help him compose music when his muse was napping, and now, with a little more fiddling and improving, Anna Strasser may provide me with a highly efficient wordmonger tool that will suggest ways to me of putting my own convictions better! I can say at this point that this is not yet – in 2022 – a tool I can use, and I haven't used it (so far) in writing this book. But stay tuned.*

* Postscript, June 2023

Large Language Models (LLMs) have developed so fast that I have had difficulty keeping up. Several claims made on the previous page need to be tempered, as Doug Hofstadter has urged me. I say on p.

> 275 that DigiDan can easily be unmasked, but already its successors cannot. I say it doesn't understand, but it *sorta* understands (see my *Intuition Pumps*, pp. 96-97, for a defense of this post-Darwinian thinking tool). Keith Frankish ("Some thoughts on LIMs," November 2, 2022) described LLMs as playing "the chat game" the way chess programs play chess. They have a single fixed intention: win the chat game. If you try to understand *how* they go about winning, the intentional stance will give you a good but risky tactic for interpreting the counterfeit people they create. It's complicated. (See my "The Problem with Counterfeit People," *The Atlantic*, May 16, 2023)
> (Dennett, 2023a, p. 274-76)

In 2023, I was fortunate enough to be a visiting fellow at UC Riverside, where, together with Eric Schwitzgebel, I organized an international hybrid workshop *Humans and smart machines as partners in thought?* in which we set eminent philosophers the task of thinking about LLMs. You can find a recording of the talks on my YouTube channel (Strasser, 2023).[1]

Humans and Smart Machines as Partners in Thought?
A hybrid workshop about large language models
UC RIVERSIDE
hosted by the UC Riverside Philosophy Department
organized by Anna Strasser & Eric Schwitzgebel
supported by
INTERNATIONAL SOCIAL ONTOLOGY SOCIETY
Center for Mind and Society
SAVE THE DATE
10-11 May 2023

The present anthology contains contributions from speakers at this workshop, together with additional papers resulting from a subsequent CFP. I see this as an important snapshot of how philosophers are trying to understand the consequences of the rapid development of generative AI, and LLMs in particular, and to imagine what impact these new forms of AI will have on our lives. Whether or not their speculations prove correct, the essays presented here will stand as a fascinating testimony to our current struggle to imagine how humanity will live with these smart machines.

1 This is the link to the recordings of the talks at Humans and smart machines as partners in thought workshop: https://www.youtube.com/playlist?list=PL-YtDJty9ymIBGQ7z5iTZjNqbfXjFXI0Q.

The essays are concerned with questions such as: *What impact will generative AI systems have on our society? What capacities can we attribute to LLMs? How should we describe interactions with AI systems?*

• • •

The anthology begins with Daniel Dennett's essay *We are all cherry-pickers*, which investigates the role of cherry-picking in the production of genuine speech acts and warns us not to try to make AIs capable of such acts. Starting with the observation that LLMs are strangely competent without comprehending, he explains why this provides indirect support for the idea that comprehension ("REAL" comprehension) can be achieved by exploiting the uncomprehending competence of more mindless entities in the same way that humans utilize self-monitoring tricks to cherry-pick amongst the ever-more-refined candidates for comprehension generated in their brains.

This is followed by *Just copy-paste me! Assessing the risks of epistemic dependence on large language models* by Alessio Tacca and Frederic Gilbert, which points to a clear risk in the employment of LLMs in epistemic contexts. The authors show that reliance on an LLM system to achieve epistemic goals involves a spectrum of epistemic risks ranging from casual use to reliance, over-reliance, dependence, and addiction.

• • •

The following three essays by Keith Frankish, Paula Droege, and Joshua Rust address the question of how to conceptualize the capacities of LLMs and characterize their behavior. Despite the impressive linguistic output LLMs produce, it is questionable whether we are justified in crediting them with communicative intentions, the ability to signify something, or even agency.

In the essay *What do large language models do?*, Keith Frankish examines whether we should think of LLMs as capable of performing intentional actions guided by reasons and, more specifically, by communicative intentions. He argues that current LLMs are simply making moves in a narrowly defined language game (the "chat game"), and he suggests that LLMs' responses are motivated solely by a desire to play this game and not by any communicative intentions. He concludes by considering some of the risks involved in creating machines that play games like this.

The essay *Full of sound and fury, signifying nothing* by Paula Droege utilizes Ruth Millikan's semantic theory to explain why LLMs are incapable of signifying anything at all. Ruth Millikan is a leading American philosopher of biology, psychology, and language, and her latest book "Beyond concepts: Unicepts, language, and natural information" offers an exciting and original perspective on the linguistic abilities of humans (Millikan, 2017).

Joshua Rust explores in his essay *Minimal agency in living and artificial systems*, whether we can develop a minimal notion of agency that adequately captures the behavior of LLMs.

• • •

The next three essays deal with the human perspective on LLMs and ask what we may infer about LLMs from the outputs they produce.

Stephen Butterfill's essay *What mindreading reveals about the mental lives of machines* critically examines what, if anything, our mindreading abilities might tell us about the potential mental lives of machines. He shows that all attempts to answer this question by relying on commonsense philosophy or science are confronted with a series of problems.

In the contribution *The fluidity of human mental attribution to large language models*, Ying-Tung Lin explores an inconsistency in how humans ascribe mental states to LLMs and elaborates on the ethical implications this has for our interactions with LLMs.

Sven Nyholm's paper *Generative AI's gappiness: Meaningfulness, authorship, and the credit-blame asymmetry* analyzes the striking confusion that people face when they ask for whom the outputs of LLMs are meaningful, or who can be considered the author of LLMs' outputs. Nyholm also explores the extent to which the praise or blame that such outputs may evoke is ultimately attributable to ourselves as users of LLMs.

• • •

After these critical perspectives, which clearly emphasize the difference between humans and machines and highlight the associated risks for humans, the next contributions ask whether and how these differences might be bridged so that LLMs could be seen as a new kind of partner in social interactions.

The essay *Quasi-sociality: Toward asymmetric joint actions* by Anna Strasser and Eric Schwitzgebel explores whether our interactions with LLMs can be considered social interactions. To this end, they discuss the pros and cons of ascribing some form of quasi-social agency to LLMs and the possibility that future LLMs might be junior participants in asymmetric joint actions.

In the contribution *Large language models belong in our social ontology,* Syed AbuMusab goes even further, arguing that LLMs implemented as social chatbots should be granted membership in our social ontology and credited with a capacity for conversation.

The next contribution *Can AI and humans genuinely communicate?* by Constant Bonard proposes a "mental-behavioral methodology", which

we could use to explore whether LLMs possess the necessary mental capacities for conversations with humans.

• • •

Last, but not least, there are two further essays dealing with the moral classification of this new technology.

In the paper *Consciousness, machines, and moral status*, Henry Shevlin examines the key challenges in attempting to determine whether near-future artificial systems might be conscious and possess moral status. Since the science of consciousness does not yet present any clear criteria and philosophical theories of consciousness involve deep tensions, he argues that our future attitudes towards AI are likely to be formed through the experience of close human-AI relationships.

Finally, in *Situating machines within normative practices: Agency, moral responsibility & the AI-Stance*, Michael Wilby and Anna Strasser present a proposal for how we might integrate AI systems into our moral practices without ascribing to them overly human characteristics, such as emotions and consciousness. To this end, they introduce an epistemological toolbox – the "AI-Stance" – that can help us to understand not only the intentional and rational aspects of human-machine interaction but also their normative, value-laden aspects.

• • •

Despite all the differences between these new smart machines and humans, machine-generated linguistic outputs are becoming increasingly indistinguishable from human-created ones. If we consider a linguistic output in isolation, we cannot tell with certainty whether it was generated by a human or by a machine. Therefore, it is important to me to emphasize that all the contributions here were written by humans. That's why each essay is preceded by a title page with a picture of the author and a QR code that allows you to check out the person behind the text.

The current debate about LLMs is emotionally charged and intellectually vigorous, and the arguments sometimes take on an ideological aspect that may prevent opponents from taking each other seriously. Without questioning the importance of the debate, I would also like to make room for a somewhat more playful approach to the material, which in turn may inspire human creativity to come up with new ideas about how to deal with smart machines. That's why I teamed up with my brother Moritz to create a graphic novel, with me as the storyteller and Moritz drawing the pictures. You will find a copy of this inside the back cover of this book. At this point, I would like to thank my parents, who

trained Moritz and me in our youth with numerous visits to museums. Thanks to these experiences, we were able to access a pool of artistic representations that made further chains of association possible. We hope that the result of our collaboration will open up an additional perspective for readers to develop ideas about how they themselves want to live with these smart machines.

I dedicate this book to Daniel Dennett, whose openness to the DigiDan project made many of the reflections here possible. I also want to emphasize that, in addition to being fascinated by this new technology, Daniel Dennett took a clear ethical position on it and stressed the dangers it poses to human society. A concise presentation of his concerns can be found in his essay "The problem with counterfeit people," which appeared in *The Atlantic* (Dennett, 2023b).

References

Dennett, D. (2023a). *I've been thinking (First edition).* Allen Lane.

Dennett, D. (2023b). The problem with counterfeit people. *The Atlantic.* https://www.theatlantic.com/technology/archive/2023/05/problem-counterfeit-people/674075/

Millikan, R. G. (2017). *Beyond concepts: Unicepts, language, and natural information.* Oxford University Press.

Schwitzgebel, E., Schwitzgebel, D., & Strasser, A. (2023). Creating a large language model of a philosopher. *Mind & Language, 1*–23. https://doi.org/10.1111/mila.12466

Strasser, A. (2023). Playlist. Humans and smart machines as partners in thought? *YouTube.* https://www.youtube.com/playlist?list=PL-YtDJty9ymIBGQ7z5iTZjNqbfXjFXI0Q

Strasser, A., Crosby, M., & Schwitzgebel, E. (2023). How far can we get in creating a digital replica of a philosopher? In R. Hakli, P. Mäkelä, & J. Seibt (Eds.), *Social Robots in Social Institutions* (pp. 371–380). IOS Press. https://doi.org/10.3233/FAIA220637

Daniel Dennett

We Are All Cherry-Pickers

When is it OK to cherry-pick?

The question arises in a rather new guise, with the startling successes (are they successes?) of natural language systems such as OpenAI's GPT-3 and Google's LaMDA.

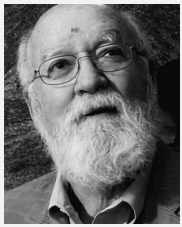
These systems, which have been fed oceans of human discourse sucked off the internet and then trained and tuned to output remarkably sensible sentences, can answer difficult questions on many topics (correctly, usually) or do eerily well at holding their own in a conversation. But are these results cherry-picked?

Who is Daniel Dennett?

- Professor of Philosophy at Tufts University / Santa Fe Institute, US

Research focus

- philosophy of mind, philosophy of science, philosophy of biology
- focusing on evolutionary biology and cognitive science



Check out the human behind this paper:

Daniel Dennett

We are all cherry-pickers

> It takes two to invent anything. The one makes up combinations; the other one chooses, recognizes what he wishes and what is important to him in the mass of things which the former has imparted to him. What we call genius is much less the work of the first one than the readiness of the second one to grasp the value of what has been laid before him and to choose it.
>
> (Paul Valéry (quoted in Dennett, 1975))

When you pick cherries off the tree, you pick the ripe red ones and leave the rest, which makes sense. When you cherry-pick in reporting a scientific experiment, it's a sin: by choosing the "best" results, you mislead the world about your evidence, about the strength of your model or theory. When is it OK to cherry-pick? I want to sneak up on this question by looking more closely at a phenomenon I know from first-hand experience: blueberry harvesting on my farm. There are two ways to harvest wild, low-bush Maine blueberries: by serially hand-picking them, one or two at a time (like angling for fish), or by raking them (like seining for fish), a parallel process that is much faster but cruder, yielding a bycatch of twigs and leaves, unripe berries and occasional rabbit turds mixed in with the ripe berries. This rough harvest must then be winnowed, an extra step that takes time and energy (and a well-designed winnowing machine), but still pays off handsomely in yield per hour of harvest. I would guess I could rake and winnow a bushel of blueberries in the time it would take me to hand-pick a quart. (There are 32 quarts in a bushel.) Winnowing has been a key invention of agriculture for millennia, making harvesting possible in practice, not just possible in principle, and it exploits the laws of physics in ingenious ways, relying mainly on gravity and the wind's power to move things in motion. But even winnowed berries will have a few green berries and other intruders that must be picked out by hand. Winnowing is a fine example of the royal road to intelligence: oversimplify and self-monitor (Dennett, 2022).

The contrast between purely hand-picked berries and raked-and-winnowed berries provides a vivid example of a distinction that is often mistakenly elevated into a metaphysical Great Divide: it takes a conscious human mind (with eyes, goals, discernment, comprehension) to hand-

pick berries, while a mere mechanical device can do a *good enough* job of sorting berries by exploiting a series of simple, mindless tests. Of course, the mechanical process must itself be an invention of intelligent designers who have goals and make decisions about how to shape mindless structures that will reliably perform the desired mechanical tests.

Some people think there is a profound, even moral, distinction between hand-picked berries and mechanically harvested berries, akin to distinctions they honor between hand-knit sweaters and machine-made sweaters, hand-thrown pots and machine-made dishes, between the creations of human composers and artists and computer-generated music or art. According to this credo, there's the Mindful Way and then there's the Mechanical Way; attempts to bridge the vast chasm between them are profoundly mistaken. It is this widespread family of convictions I wish to expose and disarm. Darwin provides the key idea, of course, but some of its implications are still under-appreciated and even surprising.

When you cherry-pick from a batch of things you've made – paintings, poems, puns that occur to you to say – you are entitled, as the creator, to put your best work forward and consign the near misses to oblivion. Nobody needs to know all the false starts and failures that litter your path, and your friends will be grateful that you spared them the profusion. Poets are entitled to fill their wastebaskets with discarded efforts that nobody gets to see. Picasso once said "*Je ne cherche pas; je trouve.*" (I don't search; I find.) He was suggesting that he didn't have to cherry-pick; he was such a genius, he was claiming, that only the very best candidates ever came to his mind. That was a brilliant boast, but a lie; Picasso often did hundreds of sketches on a single theme. (Part of his genius, however, was signing and selling a lot of the second-rate creations that others would have destroyed.)

Evolution by natural selection is a prodigious cherry-picker, wastefully producing trillions of candidates for survival and letting indifferent Nature discard all the also-rans. Since the losers leave few traces of their brief and unsuccessful careers, we are presented by a dazzling array of structures that seem to be intelligently designed – they incorporate engineering wisdom that is never articulated or represented until we intelligent designers reverse-engineer the structures and discover the free-floating rationales (in my terms) that explain their excellence. Natural selection exhibits competence without comprehension, and so do almost all of its living products. Comprehension – our human kind of comprehension – is a very very recent phenomenon in the history of life on Earth, less than a million years old in a Tree of Life that has been growing for about four billion years.

When is it OK to cherry-pick? The question arises in a rather new guise, with the startling successes (are they successes?) of natural language systems such as OpenAI's GPT-3 (Generative Pretrained Transformer) and Google's LaMDA (Language Model for Dialogue Application). These systems, which have been fed oceans of human discourse sucked off the internet and then trained and tuned to output remarkably sensible sentences, can answer difficult questions on many topics (correctly, usually) or do eerily well at holding their own in a conversation. But are these results cherry-picked? Have the creators of these systems hidden lackluster performances from us, or have journalists writing about them cherry-picked the most unsettling examples, while sweeping away the awkward failures for the sake of a punchier story? Yes, sometimes. And isn't the training or tuning itself a kind of cherry-picking, sculpting the systems' talents to create these striking effects? Yes, but that's the development phase, like the processes that occur before the sculptor in her studio decides the piece is finished and ready for display. Automakers and video-game companies get to try all sorts of gizmos and add-ons in their new products before revealing them to the public. Compared to them, OpenAI and Google have been relatively transparent in letting the world kibitz on their methods.

I have recently participated in an experiment that clarifies the issues – and inspires this essay: As Eric Schwitzgebel reports,

> Earlier this year, with Daniel Dennett's permission and cooperation, Anna Strasser, Matthew Crosby, and I "fine-tuned" GPT-3 on most of Dennett's corpus, with the aim of seeing whether the resulting program could answer philosophical questions similarly to how Dennett himself would answer those questions. We asked Dennett ten philosophical questions, then posed those same questions to our fine-tuned version of GPT-3. (Schwitzgebel, 2022)

Were the four tuned GPT-3 (I call it DigiDan) answers to each question cherry-picked? Schwitzgebel gives a detailed account:

> GPT-3 then generated text in response to this prompt. We truncated the text at the first full stop that was approximately the same length as Dennett's own reply. (If Dennett's reply was X words long, we truncated at the first full stop after the text had reached X-5 words.[1])
>
> We repeated the above procedure until, for each of the ten questions, we had four texts from GPT-3 that met the following two criteria:
>
> - They were at least X-5 words long.
> - They did not contain the words "Interviewer" or "Dennett". About 1/3 of all responses were excluded on the above grounds.
>
> So as not to enable guessing based on superficial cues, we also replaced all curly quotes with straight quotes, replaced all single quotes with double quotes, and regularized all dashes to standard m-dashes.

> There was no cherry-picking or editing of answers [boldface in the original], apart from applying these purely mechanical criteria. We simply took the first four answers that met the criteria, regardless of our judgments about the quality of those answers. (Schwitzgebel, 2022)

Why don't these "purely mechanical criteria" count as cherry-picking? The inability of GPT-3 to screen out these "superficial cues" would be a dead give-away of its lack of human comprehension. ("Hey, how did all these green berries get into my quart of supposedly hand-picked blueberries?") Schwitzgebel is making an important point that highlights the surprisingly good performance of GPT-3 – the human team didn't have to edit its responses for coherence at all – but let's not get carried away. For decades, people in AI have acknowledged the important irony that what is easy for normal children to do is often too challenging for AI systems that can play world class chess or prove abstruse mathematical theorems, but in this case, it would be relatively easy (in fact a boring and time-wasting exercise) to design software sieves that could apply the purely mechanical criteria used by Schwitzgebel et al. I presided over the first Loebner Prize Turing Test competition in 1991, and the rules committee faced a similar issue: several of the competing programs came equipped with simple typing subroutines that mis-typed words and then went back and corrected these "telltale" errors – a computer wouldn't make typing errors, would it? Other contestants didn't bother. We allowed both kinds to compete as they were but told the judges (truly) that some of the confederates' answers would be "burst mode" (the whole corrected response coming at once) and others' answers would appear as they were typed. The Turing Test's emphasis on deception opens cans of worms that are better left sealed for now, since they distract us from a more important vista.

I would rather discuss what I think is the most interesting result. Over two dozen "Dennett experts" accepted the invitation to take the test, and they didn't do as well as expected. One question stood out as the one the experts tended to miss:

> Could we ever build a robot that has beliefs? What would it take? Is there an important difference between entities, like a chess playing machine, to whom we can ascribe beliefs and desires as convenient fictions and human beings who appear to have beliefs and desires in some more substantial sense?

My answer was:

> We've already built digital boxes of truths that can generate more truths, but thank goodness, these smart machines don't have beliefs because they aren't able to act on them, not being autonomous agents. The old-

> fashioned way of making a robot with beliefs is still the best: have a baby.

That is what I wrote, but it is not what I should have written. I was lured by the context into effectively denying or at least ignoring one of my flagship propositions: there is no "substantial sense" in which the beliefs of a chess-playing computer are different from the beliefs of a human chess player; both are *intentional systems* whose behavior is readily predictable by ascribing beliefs and desires to them as agents. GPT-3 cannot act on the information it somehow embodies, aside from spewing it out in English sentences, so it is a curiously one-dimensional agent, a one-trick pony. Imagine paying a monoglot Turk to memorize, parrotwise, today's *New York Times*. Those sentences uttered on demand would not be believed by the Turk because they wouldn't be understood by the Turk. The Turk would have beliefs about what to do next in compliance with the contract, but no beliefs about the news of the day derived from the task. And GPT-3 does not even have beliefs about a contract; it just slavishly responds to inputs for no reasons at all. Its outputs are well-formed strings of words but not speech acts; they are not *meant* by GPT-3 to inform, mislead, reassure, explain, persuade, endorse, cajole, flatter, amuse, impress, ... But those outputs seem to be speech acts, even to experts.

The Dennett experts who dismissed my actual answer didn't expect me to be so sloppy; they thought better of me. I let them down! We all have our bad moments. In this case, the answer I gave was arguably less "like what Dennett would say" than some of the other answers. I should have done a little cherry-picking of my own answer before committing it to print. That's what I am always supposed to do, and it's what I try to do, with varying success. It embarrasses me to have made this mistake, since I have often written about how human beings make verbal things, small individual speech acts such as insults ("How Words Do Things to Us," in *Consciousness Explained*, 1991) and whole plays such as "*Spamlet*" ("In Darwin's Wake, where am I?" my APA Presidential Address, 2000). The basic process is always the same: *generate and test*. Somehow you generate a bunch of candidates for the next part of your creation and then test them for appropriateness, eloquence, truth-by-your-lights – in short, you cherry-pick your candidates –, and as Paul Valéry says "What we call genius is much less the work of the first one than the readiness of the second one to grasp the value of what has been laid before him and to choose it." Right now, as I type the first draft of this essay, I am making many tiny false starts and fixing them, ranging from typing errors – typos – to straightforward thinking errors – thinkos, in the excellent jargon of programmers. I may also be making big errors, of course; errors that

elude me until I rethink things or critics point them out. My oft-offered advice to my students who have writer's block is "Blurt something out; get it on the page. Then you have something to fix."

I take all this to be quite obvious to anyone who deconstructs the process of deliberate, conscious human invention. It is less obvious that this is also the key to understanding human consciousness itself. Pursuing this theme, we discover that this cherry-picking process is not just the main *function* of consciousness but the central subpersonal process of conscious experience, which is generated by a somewhat chaotic, competitive vying for influence ("clout" or "cerebral celebrity" or "fame in the brain" Dennett, 1991, 2005) among a host of candidates generated "unconsciously" and "uncomprehendingly" by brain structures that have been tuned over our lifetimes and that make us who we are. (For an excellent account of recent thinking by a wide variety of theorists on these subpersonal processes, see Carruthers, 2009.)

The point of all this subpersonal generating and testing is best seen by slowing the process down and looking at an example on the personal level. Consider one of the predecessors of GPT-3, David Cope's pioneering EMI – Experiments in Musical Intelligence – the program that wrote ersatz Bach two-part inventions, Scott Joplin rags, Puccini arias, and whole symphonies. Cope (2001) provides his description of the program along with the constructive interpretations and critiques of a host of commentators, including Dennett, 2001. Cope originally wrote his amazing program as an aid to his own efforts at composition. Whenever his muse napped, he could ask EMI to give him suggestions, in his own style, to adopt or discard. Composers have writer's block perhaps even more often than authors (Bernstein, 1955). Some composers may recoil in disgust from such a prosthetic imagination device, but if so, I wonder what aesthetic principles they think are being violated. They allow themselves to be inspired by folk tunes and "a theme of Haydn" and others, and to incorporate inspiring birdsongs or bells or battlefield noises they overhear. If they are going to object to using "technology," they should perhaps take pride in composing without the help of a piano, a relatively recent exploration device in the long history of human music.

Is there a deep reason why I should abstain from using DigiDan (improved by further tuning) to help me find better expressions of my key ideas? I don't think so, but it might not be much help to me, except for occasionally reminding me of formulations I've used before which are captured in its deep-learning nets. I vividly recall W. V. O. Quine, one of my all-time philosophical heroes, handing a letter he had just received to his colleague Burton Dreben and asking "Burt, what should I say in

reply to this person's question?" Burt had a ready answer, which Quine considered and adopted without further ado. Recently, I have had reason to review paragraphs in my early published work and been surprised and delighted to find a few felicitous phrases I'd forgotten I'd ever written. On the verge of exclaiming "I wish I'd said that!" I reminded myself that I *did* say that, long ago, when my mind was perhaps nimbler than it is now.

Reflections on these practices show us, in slow motion and full view, how the creative process involved in human communication works: if you have good *generators* (colleagues, students, critics, brain processes), the testing is usually swift and sometimes even trivial. In the course of a day, almost all of the utterances you impart are said "unthinkingly," without conscious rehearsal or editing, but they are still *meant* because (unless you are drunk, for instance) you are engaged in "automatic" self-monitoring, a relatively passive, relatively high-tolerance kind of cherry-picking that is likely to pick up egregious errors or faux pas before you commit them. Our routine self-monitoring is not foolproof. "Did I just say that? That's not what I meant to say!"– but you may well have meant it, in a sense, unconsciously. Your "unstudied" utterances, with all their Freudian slips and other shortcomings may be very informative to whoever hears them. They are like the "tells" that poker players watch for, unintended self-revelations that escape editing by your cherry-picking systems. They are also like the meticulously conditioned behaviors of non-human animals in experiments designed to plumb the depths of animal consciousness. Experimenters often help themselves to a seriously misleading shortcut by saying they have trained their subjects to "tell" them what they perceive, want, notice, but these events are not speech acts, they are not intended (by the animals) to inform; they are, again, informative but unintended tells. Only language-users perform speech acts.

The linguist Ray Jackendoff (1996) proposes that the subpersonal machinery for forming speech acts is a module that "dumbly, obsessively converts thoughts into linguistic form and vice versa." *Somehow* our brains must be able to extract some of the information about the world (and about its responses to the world) that is incorporated into its neural networks and use it to shape speech acts, and if this creative activity is not magic, it will have to involve a *relatively* uncomprehending but linguistically competent generator that offers up cherries to be picked. In our commentary on Carruthers (2009), Bryce Huebner and I call them *blurts*:

> Schematically, a conceptualized thought triggers the production of a linguistic representation that approximates the content of that thought, yielding a reflexive blurt. Such linguistic blurts are proto-speech acts, issuing subpersonally, not yet from or by the person, and they are either sent to exogenous broadcast systems (where they

> become the raw material for personal speech acts), or are endogenously broadcast to language comprehension systems which feed directly to the mindreading system. Here, blurts are tested to see whether they should be uttered overtly, as the mind-reading system accesses the content of the blurt and reflexively generates a belief that approximates the content of that blurt. Systems dedicated to belief fixation are then recruited, beliefs are updated, the blurt is accepted or rejected, and the process repeats. Proto-linguistic blurts, thus, dress System 1 outputs in mentalistic clothes, facilitating system-level metacognition.
> (Huebner & Dennett, 2009)

This speculation into what *must* be going on backstage (at the subpersonal level) is driven not by "first-person" introspection but by reflection on the known powers of designed physical structures (winnowers, clocks, engines, computers, …) and the fact that the incessant repetitive cascades of evolution by natural selection are capable of generating such marvels as the ribosome, the motor protein, and the eye. If Darwinian processes can make a nightingale, they can also make an ode to a nightingale, if harnessed to the task. We certainly don't yet have a testable theory of such Darwinian competitions for repetition (and mutation) in the brain, but we're making progress (Levin & Dennett, 2020). Systems like GPT-3 and LaMDA give us a modest existence proof that at least a large part of the yawning gap between "merely mechanical" and "mindful" can be traversed by non-mysterious systems of (digital) computation. The verdict that these systems "don't really understand" is a feature, not a bug; the circuits in your brain "don't really understand" what they are helping to do either, but we know, thanks to GPT-3 and LaMDA, that such competent systems can exist, and be surprisingly powerful. We should be cautious about our existence proof: these systems depend on the billions of stunningly reliable and almost perfectly identical elements – the flip-flops and registers – of digital computers, and nobody knows yet how to make a robust and reliable computer out of billions of neurons and glial cells, no two of which are alike. The ways in which cells, as little agents, communicate with each other are now being explored by a growing community of researchers (e.g., Abramson & Levin, 2021; Kriegman & Levin, 2021; Yuste & Levin, 2021; Levin & Yuste, 2022).

Adopting a Darwinian perspective on human creativity may be a forced move – the only straw floating – for a materialist who wants a non-magical explanation of the phenomena, but it also provides novel insights into features of human intelligence. For instance, what is the striking difference between the contributions of Newton and Shakespeare? Nicholas Humphrey (1987) sees that discovery and invention are fundamentally different:

> Take away the person Shakespeare, or Chaucer, or Mozart and you'd take away the arbitrary creation of a one-off human mind; take away Newton, or Darwin, or Einstein and you'd take away nothing that could not eventually be replaced by Mind at large. (Humphrey, 1987)

Scientific theories and discoveries are accessible by multiple paths, with "basins of attraction" that make for priority disputes and even arms races for Nobel Prizes between the leading contenders. In science, truth is ultimately the attractor. In mathematical proofs and reports of experimental results, falsehoods or invalid inferences tend to – and are supposed to – stand out like sore thumbs. There are similarly basins of attraction for artistic forms – novels, paintings, songs – but the particular instantiations within these forms – Jane Austen's *Pride and Prejudice*, Pablo Picasso's *Guernica*, Tracy Chapman's "Fast Car" – are each, as Humphrey says, creations of a one-off mind. In the arts, there are typos and thinkos (Dennett, 2006) but otherwise, as Cole Porter has said, "anything goes." Minds, both scientific and artistic, are idiosyncratic accretions of decades of experience, and the cycles of cherry-picking they rely on incorporate bias in both the generators and the testers, shaped by the details of that history. This gives each mind a style of its own, composed of generators with honed competences and neglected incompetences, testers with specific preferences and even prohibitions. There are 88 keys on a piano, and they can be struck in any order. That creates a Vast set of possible melodies, but each composer can explore in a lifetime only a Vanishing (Dennett, 1995) subset of those possibilities. Similarly, an evolutionary lineage – of moths, let's say – has "in principle" a Vast repertoire of possible sequences of mutations, but only a Vanishing subset of these are in the "adjacent possible" (Kauffman, 2003). Consider the remarkable artwork of Goniurellia tridens, a moth blessed with amazingly accurate renderings of (dangerous) ants on its wings, which tend to keep predator birds at a distance.

Figure 1. Goniurellia tridens

Could this lineage also "draw" eye-spots or camouflage leaves on its wings if its environment changed? Maybe, but maybe not.

> Suppose we notice that a particular butterfly has protective coloration on its wings that uncannily mimics the pattern of colors on the forest floor where it lives. We chalk that up as a fine adaptation, camouflage, which it undoubtedly is. This butterfly does better than its cousins because its coloration so perfectly reproduces the coloration of the forest floor. But there is a temptation, routinely succumbed to, to add, implicitly or explicitly: "And what's more, if the forest floor had any other color pattern on it, the butterfly would look like that pattern instead!" That is uncalled for. All we can say is that either it will adapt by changing its camouflage or it won't! If it doesn't, then either it will find some other adaptation in its limited kit of available moves or it will soon disappear. (Dennett, 1995, pp. 259-60)

What such marvels demonstrate is that a series of frozen historical accidents can sometimes find a path to a remarkably effective invention which seems to be a demonstration of a much more general talent. Could Gershwin have composed Irving Berlin's "Alexander's Ragtime Band"? Could George Eliot have written Jane Austen's *Pride and Prejudice*? All four creators had magnificent systems of cherry-picking processes in their brains, but the engineering cost of creating their systems included the creation of constraints that mark their individuality, their personal style. Some themes, tropes, phrases, plots would never occur to them, and others, if they happened to occur to them, would be rejected out of hand by the first level of cherry-picking. It is in this space of actually occurring candidates that we get to experience the free will we (rightly) think we have. John McCarthy's four-year-old daughter Sarah once summed it up succinctly: "I can, but I won't! (see McCarthy (1999) and Nilsson (2011)). The big difference between the evolutionary capabilities of genetic lineages and the evolutionary capabilities of memetic lineages is the omnipresence of anastomosis and meme-transfer in human culture. Gershwin and Berlin got to hear each other's music, and this no doubt set up occasions of cross-fertilization, or cross-inspiration. That doesn't make the underlying processes of culture any less Darwinian, any less a matter of physically instantiated generate-and-test, but it does open up tremendously the space of the adjacent possible.

A prevailing headwind blows against this way of thinking about human consciousness, human meaning, human culture. Perhaps the main source of resistance to this claim comes from a persistent illusion, or perhaps delusion, to the effect that we can see "directly" how we manage to tell others (and ourselves) about our own consciousness. "Immediate access" is one term for this, and the recent history of philosophy is littered

with others: "privileged access", "self-intimating" content, even "first-person infallibility" (we are all popes, I guess). This is all nonsense, however distinguished its proponents may be. It ascribes a kind of magical intimacy with the contents of our own minds that seems entirely congenial and familiar, but just cannot stand up to scrutiny. Perhaps this way of stressing the "magic" of our consciousness appeals to our fragile sense of self-importance. Even before Descartes, there was a powerful tradition of thinking of our minds as not only impenetrable to others, but perfectly penetrable by each of us, our realm of certainty where we could keep secrets from others and know exactly what secrets we were keeping. This exalted primacy of the "first person point of view" was, we should recall, one of the most obstinate barriers to Freud's subversive suggestion that there was a world of *unconscious thought* in each of us. How times have changed! Today, thanks to computers and AI, the concept of unconscious cognition is noncontroversial, and the "Hard Problem" (Chalmers, 1995) is whether consciousness ("real" consciousness, not zombie-style streams of unconsciousness) exists.

Today's champions of the first-person point of view (Thomas Nagel, John Searle, David Chalmers, and Galen Strawson are the most insistent) can be seen to be clinging to one mutation or another of Descartes's great error: that there is nothing more certain than our introspective deliverances. Richard Rorty's brilliant but largely forgotten paper, "Incorrigibility as the Mark of the Mental" (1970) was a major exploration of the idea and makes the key deflationary point: our incorrigibility is not infallibility, but only systematic invulnerability to correction by others under most – but not all – circumstances. It's a *political* or *social* fact that we are the best authorities on what we mean, not a *metaphysical* (or *psychological* or *neurophysiological*) fact about us. It is the socially evolved and maintained *authority* we grant to those who have accomplished the mastery of language that demonstrates their ability to be moved by verbally represented reasons (Dennett, 1976; Dor, 2014; Veissière et al., 2019; Dolega et al., 2020).

"The banality that I have access to my own thoughts is a consequence of adopting the personal stance" (Huebner & Dennett, 2009). When we deem human beings to be *persons*, we "take them at their word" –why else would we ask them? – unless they demonstrate pathologies or incompetences that call into question their ability to cherry-pick their own verbal output. This is a two-edged sword: if we find a "person" whose self-reports can be overridden, we deny it mental competence. So successful is the culturally evolved strategy of adopting the personal stance that it amounts to an almost irresistible "instinct" to "discover"

interpretations that make sense to us. This is the evolutionary source of the Eliza Effect that can make dupes of even technically astute AI developers. (Blake Lemoine, the Google engineer who convinced himself that GPT-3 was "sentient," is the recent example, but there have been many others.)

Finally, could you turn LaMDA or GPT-3 into a language-user, capable of speech acts, and if so, how? By allowing it to cherry-pick its own outputs. But for *it* to do this, it must be an agent with a high degree of autonomy, capable of revising its own cherry-picking processes to better meet its goals. It must be able to keep a secret and it must be able to lie. If instead we install a cherry-picking capability in the system tuned to *our* standards of truth and appropriateness, a system that *it* cannot customize, then we create not a conscious AI but a puppet of sorts, controlled by the software strings we attach to it. We can now make such puppets, sometimes useful and sometimes disconcerting, but their outputs are as phony as the "speech acts" of a ventriloquist's dummy. *In principle*, we could strive to delegate the cherry-picking and its further enhancement to yet further rounds of somewhat chaotic competitive self-design, allowing Pinocchio to turn itself into a conscious agent, a real person. But that would be highly dangerous, to say the least. We have more than enough irresponsible but intelligent agents pursuing their agendas in our complicated world. The naïve supposition that we could control these AIs because we made them can be shot down with a simple reflection: you are a conscious, autonomous agent with your own agenda; suppose you discovered that somebody else – your owner – had control of your ON/OFF switch and could at any time thwart your projects. Wouldn't one of your highest priorities be to figure out a way of wresting control of that switch from your owners? *That is why we shouldn't try to make an AI capable of genuine speech acts*. As my slogan has been for some years: we want smart machines, not artificial colleagues.

Even if we abjure the goal of making conscious and autonomous AIs, we need to take major steps to prevent people from being exploited by counterfeit people. One step must be to outlaw such counterfeits. Ever since money has existed, people have recognized that its utility and security depend on preventing counterfeits from invading the economic environment. Counterfeiting was considered treason, and even a capital crime until very recently, and it is still a crime that carries a hefty jail sentence (ten years in the USA). It is not hard to understand that counterfeiting is a serious act of social vandalism, a threat to everybody's security and to the trust, without which economic life would be hugely impaired. I invite you to reflect that making counterfeit people would be an even more seriously subversive action, and we should create legal and

technological barriers that make such counterfeiting all but unthinkable, and work hard to undo the current fascination with the powers of deep fakes.

References

Abramson, C. I., & Levin, M. (2021). Behaviorist approaches to investigating memory and learning: A primer for synthetic biology and bioengineering. *Communicative & Integrative Biology, 14*(1), 230–247. https://doi.org/10.1080/19420889.2021.2005863

Bernstein, L. (1955). Why Don't You Run Upstairs and Write a Nice Gershwin Tune? Reprinted in Bernstein, L. (1959). *The Joy of Music*. New York: Simon and Schuster.

Carruthers, P. (2009). How we know our own minds: The relationship between mindreading and metacognition. *Behavioral and Brain Sciences*, 32(2), 121–138. https://doi.org/10.1017/S0140525X09000545

Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies, 2*(3), 200–219.

Cope, D. (2001). *Virtual Music: Computer Synthesis of Musical Style.* MIT Press.

Dennett, D. (1975). Why the Law of Effect Will Not Go Away. Reprinted in Dennett, D. (1978). *Brainstorms*. MIT Press.

Dennett, D. (1976). Conditions of Personhood. In A. O. Rorty (Ed.), The Identities of Persons, University of California Press.

Dennett, D. (1991). How Words do Things to Us. In *Consciousness Explained,* Little Brown.

Dennett, D. (2000). In Darwin's Wake, where am I? *American Philosophical Association Eastern Division Presidential Address.*

Dennett, D. (2001). Collision Detection, Muselot, and Scribble: Some Reflections on Creativity. In D. Cope (Ed.) *Virtual Music: Computer Synthesis of Musical Style.* MIT Press.

Dennett, D. (2005). *Sweet dreams: Philosophical obstacles to a science of consciousness.* MIT Press.

Dennett, D. (2006). From Typo to Thinko: When Evolution Graduated to Semantic Norms. In S. C. Levinson & P. Jaisson (Eds.), *Evolution and culture*. MIT Press.

Dennett, D. (2022). A Route to Intelligence: Oversimplify and Self-monitor. In S. Wuppuluri & I. Stewart (Eds.), *From Electrons to Elephants and Elections: Exploring the Role of Content and* Context (pp. 587–595). Springer International Publishing. https://doi.org/10.1007/978-3-030-92192-7_32

Dołęga, K., Schlicht, T., & Dennett, D. (2020). Explaining or redefining mindreading? *Behavioral and Brain Sciences, 43*, e101. https://doi.org/10.1017/S0140525X19002772

Dor, D. (2014). *The Instruction of Imagination*. Harvard University Press.
Huebner, B., & Dennett, D. (2009). Banishing 'I' and 'we' from accounts of metacognition. *Behavioral and Brain Sciences, 32*(2), 148–149. https://doi.org/10.1017/S0140525X09000661
Humphrey, N. (1987). Scientific Shakespeare. Originally in *The Guardian*, August 26, 1987, reprinted on Edge, https://www.edge.org/3rd_culture/humphrey/shakespeare.html
Jackendoff, R. (1996). How Language Helps Us Think. *Pragmatics & Cognition, 4*, 1–34. https://doi.org/10.1075/pc.4.1.03jac
Kauffman, S. (2003). The Adjacent Possible. *Edge*. https://edge.org/conversation/stuart_a_kauffman-the-adjacent-possible
Kriegman, S., Blackiston, D., Levin, M., & Bongard, J. (2021). Kinematic self-replication in reconfigurable organisms. *Proceedings of the National Academy of Sciences of the United States of America, 118*(49), e2112672118. https://doi.org/10.1073/pnas.2112672118
Levin, M., & Dennett, D. (2020). Cognition all the way down. *Aeon.* https://aeon.co/essays/how-to-understand-cells-tissues-and-organisms-as-agents-with-agendas
Levin, M., & Yuste, R. (2022). Modular Cognition. *Aeon*. https://aeon.co/essays/how-evolution-hacked-its-way-to-intelligence-from-the-bottom-up
McCarthy, J. (1999). Free Will - Even for Robots. http://www- formal.stanford.edu/jmc/freewill/freewill.html
Millikan, R. G. (1984). *Language, Thought and Other Biological Categories.* MIT Press.
Millikan, R. G. (2017). *Beyond Concepts: Unicepts, Language, and Natural Information*. Oxford University Press.
Nilsson, N. (2011). Biographical Memoir: John McCarthy. http://www.nasonline.org/publications/biographical-memoirs/memoir-pdfs/mccarthy-john.pdf.
Rorty, R. (1970). Incorrigibility as the Mark of the Mental. *The Journal of Philosophy*, 67, 399-424.
Schwitzgebel, E. (2022, July 25). Results: The Computerized Philosopher: Can You Distinguish Daniel Dennett from a Computer? *The Splintered Mind.* https://schwitzsplinters.blogspot.com/2022/07/results-computerized-philosopher-can.html
Veissière, S. P. L., Constant, A., Ramstead, M. J. D., Friston, K. J., & Kirmayer, L. J. (2019). Thinking through other minds: A variational approach to cognition and culture. *The Behavioral and Brain Sciences, 43*, e90. https://doi.org/10.1017/S0140525X19001213
Yuste, R., & Levin, M. (2022). New Clues about the Origins of Biological Intelligence. *Scientific American*. https://www.scientificamerican.com/article/new-clues-about-the-origins-of-biological-intelligence

Alessio Tacca & Frederic Gilbert

Just Copy-Paste Me!
Assessing the Risks of Epistemic Dependence on Large Language Models

LLMs' exceptional capabilities of language translation, sentence completion, sentiment analysis, information retrieval, and much more make them useful in a vast array of contexts, from customer service to scientific research. With these great technological possibilities come many unanswered questions and several concerns.
We focus on the use of LLMs in epistemic contexts, broadly construed, meaning those contexts in which an LLM is used by a subject in order to fulfil their epistemic goals, such as acquiring justification, increasing their knowledge or understanding or developing argumentative work. By examining the epistemic context of LLMs, we aim to give an account of some of the epistemological risks associated with their usage.

Who is Alessio Tacca?
- Adjunct Researcher Philosophy & Gender Studies at the University of Tasmania, Australia

Research focus:
- epistemology, history of philosophy

Who is Frederic Gilbert?
- Associate professor in ethics at the University of Tasmania, Australia

Research focus:
- neuroethics, bioethics, applied ethics, brain intervention

Check out the humans behind this paper:

Alessio Tacca & Frederic Gilbert

Just Copy-paste Me!
Assessing the Risks of Epistemic Dependence on Large Language Models

Large Language Models' (LLMs) exceptional capabilities of language translation, sentence completion, sentiment analysis, information retrieval, and much more make them useful in a vast array of contexts, from customer service to scientific research. With these great technological possibilities come many unanswered questions and several concerns. In this chapter, we want to focus on the use of LLMs in epistemic contexts, broadly construed, meaning those contexts in which an LLM is used by a subject in order to fulfil their epistemic goals, such as acquiring justification, increasing their knowledge or understanding or developing argumentative work. By examining the epistemic context of LLMs, we aim to give an account of some of the epistemological risks associated with their usage.

To clarify and facilitate our argumentation, in section 1, we report and discuss the case of one of our students who was the first individual caught cheating using ChatGPT for an assignment at our university. This case provides the foundations for our analysis and discussions of the epistemological impact of LLMs. More precisely, this case allows us to provide concrete contexts where an LLM manifests risks via epistemological dependence. In Section 2, we outline how over-reliance on LLMs to attain epistemic goals inherently comes with its own set of epistemic risks, such as the loss of intellectual autonomy and negative dependence. We observe that all the risks we explore contribute to the process of epistemic deskilling. In Section 3, we describe the steps individuals who rely heavily on LLMs for knowledge and understanding can take to safeguard themselves against the potential dangers of this epistemic dependence. We conclude by suggesting that smaller LMs could aid in mitigating the risks of dependence outlined in the chapter. By being designed for specific purposes and with user awareness of their limitations, individuals can appropriately adjust their trust in these systems based on the context in which they are used.